

# Detection of construction biases in biological databases: the case of miRBase

Guilherme Bicalho Saturnino,<sup>1,\*</sup> Caio Padoan de Sá Godinho,<sup>2,†</sup> Denise

Fagundes-Lima,<sup>1,‡</sup> Alcides Castro e Silva,<sup>3,§</sup> and Gerald Weber<sup>1,¶</sup>

<sup>1</sup>*Departamento de Física, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil*

<sup>2</sup>*Departamento de Ciências Biológicas, Universidade Federal de Ouro Preto, Ouro Preto, MG, Brazil*

<sup>3</sup>*Departamento de Física, Universidade Federal de de Ouro Preto, Ouro Preto, MG, Brazil*

Biological databases can be analysed as a complex network which may reveal some its underlying biological mechanisms. Frequently, such databases are identified as scale-free networks or as hierarchical networks depending on connectivity distributions or clustering coefficients. Since these databases do grow over time, one would expect that their network topology may undergo some changes. Here, we analysed the historical versions of miRBase, a database of microRNAs where we performed an alignment of all mature and precursor miRNAs and calculated a pairwise similarity index. We found that the clustering coefficient shows important changes during the growth of this database. For two consecutive versions of the year 2009 we found a strong modification of the network topology which we were able to associate to a technological change in miRNA discovery. To evaluate if these changes could have happened by chance, we performed a set of simulations of the database growth by sampling the final version of miRBase and creating several alternative histories of miRBase. None of the simulations were close to the actual historical evolution of this database, which we understand as a clear indication of a very strong construction bias.

Since Barabasi's seminal work [1] the concept of complex networks has spread to nearly all fields of research. Countless systems can be modelled as complex networks, for instance in social interactions (human relationship, inter-city and intra-city movements [2]), ecology (predation, mutualism [3, 4]), transport (traffic flow, airports [5]). Even food recipes can be analysed in this way [6]. Unsurprisingly, networks built from interactions in molecular biology were among the first applications of complex network analysis and have since become a standard part of bioinformatics textbooks, see for example Ref. 7. One of the best known examples of networks in molecular biology are protein interaction networks where the network is formed with proteins that bind to each other in order to carry out some biological function [8, 9]. Also, gene regulatory networks where genes interact via transcription factors and the metabolic network [10] are examples of network topology analysis. The usage of biological networks has become so common that even a word was invented to describe more synthetically the whole set of interactions in a cell: the interactome.

The growing popularity of network topology analysis in molecular biology is easily explained in terms of the promise to uncover fundamental new biological insights while analysing massive amounts of data. Yet one key question appears to have been systematically overlooked: what if the network topology is not of biological origin but the consequence of particular technological constraints by which the database was populated with biological data? To stay within the example of well known

protein interaction networks, suppose that some network analysis shows that a given protein is found to form a hub of highly connected interactions. One possibility would be that this particular protein simply attracted more research activity resulting in more entries in the database while other proteins received less attention. Clearly, network topology analysis could be entirely skewed by biases in the construction of the database. To our knowledge, this problem has not seen attention or being acknowledged. Note that we are not discussing in this work the specific issue of scale-free topologies of biological networks which were found to be problematic [11–13].

The main problem here is that biological databases are inherently, but not uniformly, incomplete. The present day database is essentially a subset of what biology has in store. Therefore database construction biases can be regarded as a problem of subnet sampling.

Sampling a part of a large network is a common and mostly unavoidable procedure especially for very large networks. For example it is impossible to analyse the network topology of the entire internet since many network quantities scale to the square or even to the cube of the number of nodes. This problem results in the following classical question: does the subnet share the same topological properties of the whole net? Evidently, this is a difficult question which attracted some attention, sometimes with seemingly contradicting findings [14, 15]. One definite conclusion though is that the resulting topology is crucially dependent on the sampling strategy and that the property of clustering appears to be the most sensitive to sampling [16].

For the question we wish to analyse, if there is a database construction bias, there is no choice of sampling strategy available. The subnet sampling is entirely determined by whatever data was obtainable by current technical methods at a certain period of time. Consider for example, a network which grows over time. Any pre-

\* guilherme.bicalho2@gmail.com

† godinhocps@gmail.com

‡ defalima@gmail.com

§ alcidescs@gmail.com

¶ gweberbh@gmail.com

vious version of this network is a subnet of the newer version. Clearly, in this case the network topology could change over time. For instance, a network could start out as random in its early days and gradually mature into a scale free network if some preferential attachment is involved.

The question therefore is not only if there is a construction bias at all, but how much data is necessary for this bias to become negligible. One aspect of database formation makes this question hard to answer: there are many levels of information in a database and all of them depend to some extent on annotation. For example, a database may contain a specific genomic sequence and also if this sequence is related in some way to other sequences. To form a network one usually combines multiple informations to connect different nodes which means that the network topology results from an entangled mix of information.

Here we will attempt to address the detection of database construction biases by using the smallest amount of information possible. To achieve this we selected one relatively small yet important database: the database of microRNA (miRNA) known as miRBase [17]. The discovery of miRNAs is fairly recent and the database has seen a peak of activities over the last years [18]. Also, the discovery of new miRNAs has improved considerably since the early days of this database. We can now ask our question again for this database: is the network topology of this database biased by the way new miRNAs are discovered? If yes, which topology parameters would be affected and how strongly?

To form a network of miRNAs we need to choose how to link them. There are many possibilities, for example we could link them by their silencing targets, by the genes which express them or a mixture of different criteria. The problem with this approach is that the link would depend on multiple layers of database annotation and would be a source of complication for our analysis. Instead, we selected to establish this relation with sequence similarities which has the advantage of being a totally deterministic way of establishing a relation between two miRNAs. Therefore, our network depends solely on one information contained in the database which is the miRNA sequence and nothing else. Another motivation is that some of the methods for discovering new miRNAs rely on sequence similarities and therefore this may become apparent in the network topology.

A further reason which makes miRNAs attractive for similarity network analysis is their short length, around 20 nt for mature miRNAs and of the order of 100 nt for precursor sequences. It is their short lengths which allowed us to calculate the pairwise sequences similarities for the complete miRBase. We calculated the sequence similarities of all miRNA sequences compared to each other which provides us with a similarity index. Then we calculated the network topology parameters such as average weights and clustering coefficients for each of the historical versions of miRBase. We performed a series

of simulations which basically ask the following question: what if history had happened differently? What if the miRNAs were discovered in a different order? Can we actually distinguish the real database from the simulated one? By analysing a subset of the final database we are essentially borrowing robust statistics or bootstrapping methods. It is common to choose randomly a subset of a larger set of data and then estimate how much a given parameter deviates when compared to the original set [19]. For instance, in a recent work [20] we were able to estimate error bars for DNA flexibility using this type of method.

The main rationale behind our approach is that if sequences were deposited in a database in a certain biased order any resulting network topology, no matter which linking criteria was used to form the network, will be biased as well. Note that an unbiased database should result in similar network topology parameters over all database versions. All topology parameters should confirm this. If just one parameter presents a bias this establishes unambiguously that the dataset history is biased and no further tests are necessary. Considering that similarity networks also have found their uses for sequence clustering [21], for multiple alignment of protein sequences [22], phylogenetic analysis [23] and visualization of protein superfamilies [24], the conclusions presented here may be relevant for these applications.

## METHODS

### Similarity index

We used a standard Needleman-Wunsch (NW) [25] alignment algorithm to calculate the sequence similarity of all miRNAs against each other. The alignment matrix  $P$  is filled according to the following rule

$$P(i, j) = \max \begin{cases} P(i-1, j-1) + R(i, j) \\ P(i-1, j) + g \\ P(i, j-1) + g \end{cases} \quad (1)$$

where  $i$  and  $j$  are the nucleotide positions of the two sequences which are being compared, and

$$R(i, j) = \begin{cases} m & \text{if } i \text{ and } j \text{ are the same nucleotide} \\ d & \text{otherwise} \end{cases} \quad (2)$$

The alignment matrix  $P$  expresses the degree of similarity between two sequences. Each cell of the matrix is evaluated sequentially using the rule  $R$  and compared against the previous cells. If the two sequences are very similar the score  $m$  is used more frequently producing higher values for these cells. In this work we used the score  $m = 2$  which is used when two nucleotides are identical. The score  $d = -1$  is a penalty when two nucleotides are different, and gaps in the sequences are penalized even more by  $g = -2$ . For sequences of length  $l$  the highest score  $S$  that can be achieved is  $m^l$ , that is, the score  $m$

( $l$  nucleotides are identical) is summed over  $l$  times when evaluating Eq. (1). We then normalize the NW score to one, which we call the similarity index  $s = S/m^l$ . Therefore, two identical sequences will score  $s = 1$ .

These similarity indices are used to build a weighed network where  $w_{ij}$  is the similarity  $s$  between miRNAs  $i$  and  $j$ . One advantage of using the NW algorithm is that it results in symmetrical weights, that is  $w_{ij} = w_{ji}$ , therefore avoiding problems with unsymmetrical similarities [23] and other errors associated to Blast searches [26]. For  $N$  microRNAs this results into  $N(N-1)/2$  similarity relations and to reduce the size of the networks we remove all links below a minimal score  $w_{ij} < s_m$  [27]. We have chosen a minimal score  $s_m = 0.4$  that was found to retain all nodes with at least one connection. This removal is also necessary since similarity indices below  $s \approx 0.25$  may occur by chance and therefore have no real similarity meaning (see also Fig. 1 of Ref. 27).

### Clustering coefficient

The discrete clustering coefficient is commonly defined as

$$c_i = \frac{1}{k_i(k_i - 1)} \sum_{j,h \neq i} a_{ij}a_{ih}a_{jh} \quad (3)$$

where the coefficients  $a_{ij}$  will be equal to one if there is a connection between the nodes  $i$  and  $j$  and zero otherwise. The connectivity, that is, the total number of connections leading to node  $i$  is  $k_i$ . Note that in the similarity network all nodes are connected by a weight  $w_{ij}$ , therefore the discrete clustering coefficient only differs from unity if we apply a minimal score as described in the previous section.

The weighted clustering coefficient which is formally closest to the discrete coefficient of Eq. 3 is based on the geometric average proposed by Onnela et al. [28]

$$c_i^w = \frac{1}{k_i(k_i - 1)} \sum_{j,h=1}^N (w_{ij}w_{ih}w_{jh})^{1/3} \quad (4)$$

where  $w_{ij}$  is the normalized weight of the connection between the nodes  $i$  and  $j$ , and  $k_i$  is the connectivity of node  $i$ . Basically this coefficient counts how many triplets a node and its neighbours can form. The coefficient is normalized such that it will be 0 if there are no triplets and 1 if all triplets are maximum. The clustering coefficient is a node-related property and it is common to define a global discrete and weighted clustering coefficient, averaging over all  $c_i^w$ , namely:

$$C = \frac{1}{N} \sum_{i=1}^N c_i \quad (5)$$

$$C^w = \frac{1}{N} \sum_{i=1}^N c_i^w \quad (6)$$

### Simulation

We performed several simulations to determine the database topology if the miRNAs deposited in miRBase were discovered in a different chronological order. Two different types of alternative first versions were attempted, either we sampled randomly from the last version or we started with the same version as the initial database 1.2. The simulations described in this section were performed for both mature and precursor miRNAs.

*a. Simulation with fixed initial version (FIV)* We selected all sequences in the initial version 1.2 that do exist up to version (17.0), that is, the original version 1.2 minus the sequences which were deleted up to the last version. In this case, every time we simulate new higher versions we start over with exactly the same initial version.

*b. Simulation with random initial version (RIV)* Here, the new initial version corresponding to database 1.2 is randomly populated with sequences from version 17 until we reach the same number of elements as the original version 1.2.

*c. Populating the remaining database versions* The versions after 1.2 of the database are populated randomly and without repetition, sampling from version 17.0 with the following criteria

1. first we make a set of sequences in 17.0 which have a similarity index  $s > 0.95$  with any sequence already present in the database. We then sample uniformly from this set.
2. should we run out of sequences with  $s > 0.95$  before completing the new version, then we sample randomly the remaining sequences from version 17.0.

Once a given version is completely populated and contains the same amount of sequences as the original version we perform all network topology calculations. For each database version this simulation is carried out 20 times.

This strategy intends to simulate a process by which the next version of the database is populated by sequences which are closely related to sequences which already exist in previous versions. In this way we try to mimic the effect that new sequences may be found by similarity to known ones, that is we try to introduce some level of bias.

*d. Simulation for mature *H. sapiens* miRNAs (HS)* Here we select only the mature miRNAs for *H. sapiens* which is the largest subset of miRBase and rebuild the versions containing the same of *H. sapiens* in each version.

### Sequences used

We used all mature and precursor sequences and versions of miRBase up to version 17.0 corresponding to the

TABLE I. Summary of the miRBase database versions. Shown are the sequential index, the official version designation, the date on which this version was released, the number of precursor miRNAs  $N_{\text{pre}}$ , the number of mature miRNAs  $N_{\text{mat}}$ , the number of mature miRNAs which are not present in the last version  $R_{\text{mat}}^{17}$ , and the number of mature miRNAs of *H. sapiens*  $N_{\text{HS}}$ .

index	version	date	species	$N_{\text{pre}}$	$N_{\text{mat}}$	$R_{\text{mat}}^{17}$	$N_{\text{HS}}$
1	1.1	01/2003	5	265	265	112	87
2	1.2	04/2003	5	295	256	103	87
3	1.3	05/2003	5	332	295	103	87
4	1.5	07/2003	5	400	370	160	88
5	2.0	07/2003	6	506	464	171	135
6	2.2	11/2003	7	593	528	188	136
7	3.0	01/2004	8	719	644	212	152
8	3.1	04/2004	8	889	807	242	169
9	4.0	07/2004	10	1185	1143	279	188
10	5.0	09/2004	12	1345	1298	288	189
11	5.1	12/2004	13	1420	1359	257	207
12	6.0	04/2005	21	1650	1591	378	211
13	7.0	06/2005	33	2909	2634	348	313
14	7.1	10/2005	37	3424	3102	366	319
15	8.0	02/2006	37	3518	3229	368	328
16	8.1	05/2006	39	3963	3685	408	455
17	8.2	07/2006	39	4039	3834	418	454
18	9.0	10/2006	43	4361	4167	478	470
19	9.1	02/2007	43	4449	4274	476	470
20	9.2	05/2007	46	4584	4430	494	471
21	10.0	08/2007	49	5071	4922	284	555
22	10.1	12/2007	56	5395	5234	326	564
23	11.0	01/2008	62	6396	6211	356	677
24	12.0	09/2008	77	8619	8273	408	697
25	13.0	03/2009	94	9539	9169	504	703
26	14.0	09/2009	103	10867	10566	508	718
27	15.0	04/2010	120	14197	15632	641	1100
28	16.0	08/2010	129	15172	17341	592	1223
29	17.0	04/2011	138	16772	19724		1733

version when this work was started [17, 18]. Table I summarizes all sequences used in this work.

## RESULTS AND DISCUSSION

In Fig. 1 we show the average clustering weighted coefficient  $C^w$  of mature miRNAs for each version of miRBase. During the initial years the clustering coefficients (red boxes) grow steadily reaching a peak in 2005 from which on they start decreasing again. In 2009 we observe a dramatic reduction in clustering coefficients which occurs between versions 13 and 14. We simulated the alternative history of the database with two different starting

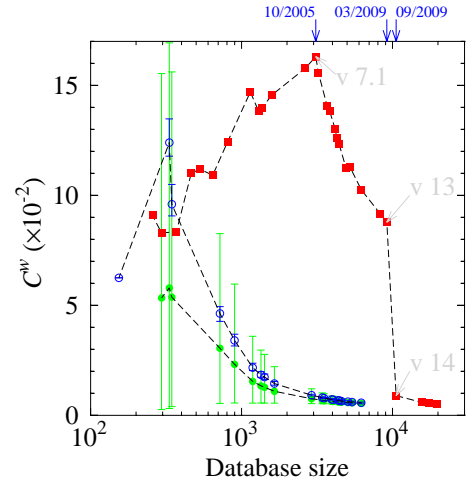


FIG. 1. **Average weighted clustering coefficients  $C^w$  of mature miRNAs as a function of database size.** Red boxes are for the real miRBase versions and blue/green circles are for the simulated database versions. The simulated database versions are represented by blue circles for a random initial database, and green boxes if they start from version 1.2.

versions. In one (blue circles) we started with an entirely random initial version (RIV) which is sampled from the last version 17. The other one (green boxes) starts with the real version 1.2 (FIV), but then grows randomly from one version to the next. Neither simulation procedure comes close to reproducing the clustering evolution of the real database, despite the strong bias introduced in the sampling procedure (see Methods). In both cases, the clustering coefficients are rapidly dominated by the low connectivity of the miRNAs which appear since version 14. From version 13 to the last 17, the database nearly doubles in size which means that at least half of the miRNAs are sequences with very low clustering coefficient. This would explain the difficulty in obtaining higher clustering coefficients in any database versions.

Fig. 2a shows the difference in clustering distribution between versions 13 and 14. While the clustering coefficients are distributed over an extended range for version 13, for the next version they form a delta-like distribution at a very low value. The distribution of weighted clustering coefficients  $C^w$  as function of connectivity  $k$ , or  $C^w(k)$ , shown in Fig. 2b confirms this sudden and dramatic change in network topology. For version 13, the average clustering coefficients are roughly constants in the log-log plot until  $k = 400$  ( $\log k = 2.6$ ). In contrast, for version 14, the clustering coefficients shown in Fig. 2b follow a power law which resembles that of a hierarchical network until  $k = 200$  ( $\log k = 2.3$ ). The connectivity distribution  $P(k)$  however, is closer to a random network, as shown in Fig. 3. The inset of Fig. 3, suggest a slight deviation from a purely random network.

One question which arises is if this change between

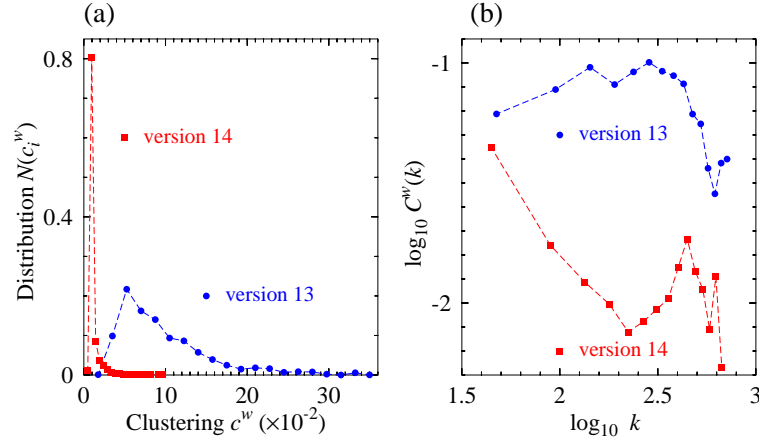


FIG. 2. **Clustering coefficient distribution of mature miRNAs.** Shown are (a) the histograms  $N(c_i^w)$  and (b) the average clustering coefficients as function of connectivity  $C^w(k)$ . Blue circles are for version 13 and red solid boxes for version 14.

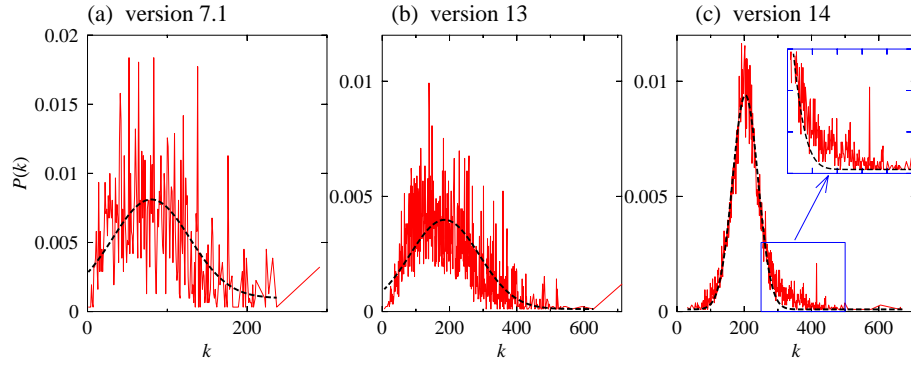


FIG. 3. **Connectivity distribution  $P(k)$  as a function of connections  $k$  of mature miRNAs.** Shown are versions (a) 7.1, (b) 13 and (c) 14. Solid red curves are the  $P(k)$  distributions and dashed black curves are the calculated symmetrical Gaussian regression of  $P(k)$ .

versions 13 and 14 is driven by variations in similarity indexes which represent the weight in our network. In Fig. 4 we show the change of average weight  $\langle w \rangle$  with database size (right scale). Overall we see a very similar behaviour as for the global clustering coefficients  $C^w$  with the notable exception of the sudden drop between versions 13 and 14. Fig. 4 also shows the global discrete clustering coefficients  $C^k$  (left scale), that is, where the weights were replaced by 1 or 0 depending on the miRNAs being connected or not with a score larger than 0.4. In this case we see the sudden decrease, which leaves us with the conclusion that it is entirely due to a complete change in network connectivity. In other words, the new miRNAs deposited in version 14 of miRBase compare to each other with basically the same similarity index as in version 13, but they do so with much fewer other miRNAs.

Precursor miRNAs, that is the longer and unprocessed sequences which results in mature miRNAs, display a very similar evolution of clustering coefficients shown in

Fig. 5. We observe the same peak for version 7.2 as for mature miRNAs, and also the sudden drop between versions 13 and 14. Overall, the clustering coefficients are lower for the pre-miRNAs than for the mature miRNAs, which is mainly due to the fact that similarity indexes tend to be much smaller between miRNAs. Smaller indexes are expected for longer sequences where it becomes less probable that two sequences are similar by chance. Nevertheless, the sudden reduction in Fig. 5 is unmistakable. As for mature miRNA, the simulations do not reproduce larger clustering coefficients and are rapidly dominated by the less connected pre-miRNA of later versions of miRBase. For the simulation which starts with the real initial version 1.2 (FIV), the subsequent versions display some large clustering coefficients which do not decrease so rapidly. The simulation with initial random database version (RIV) basically stays constant at the same average clustering coefficient for all versions. Only the initial database version shows a very large error bar with nearly disappears for the next versions.

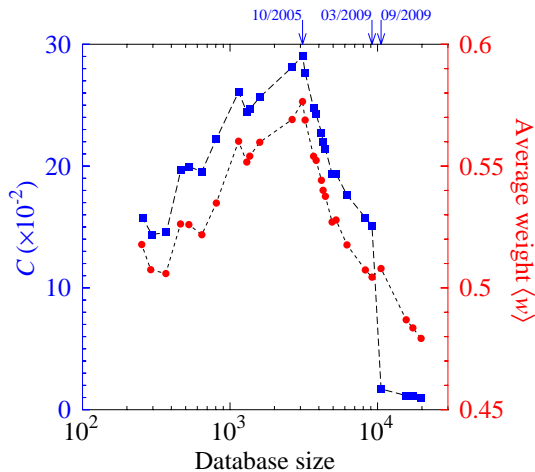


FIG. 4. Average weights  $\langle w \rangle$  and discrete global clustering coefficient  $C$  of mature miRNAs as a function of database size. Blue boxes are the average discrete clustering coefficient (left scale) and red bullets are the average weights (right scale).

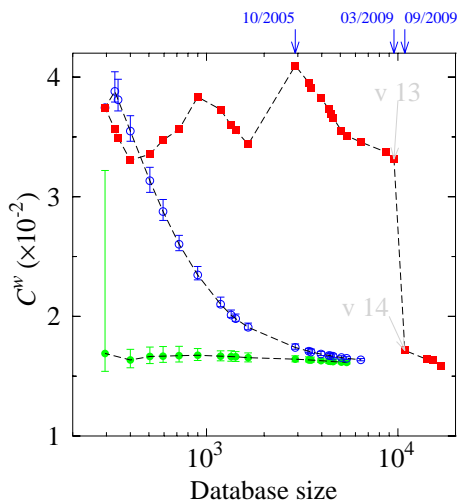


FIG. 5. Global clustering coefficients  $C^w$  of pre-miRNAs as a function of database size. Red solid boxes are for the real miRBase versions. The simulated database versions are represented by blue circles for a random initial database, and green boxes if the they start from version 1.2.

Human miRNAs is the most numerous species group present in miRBase, making up about 10% of the final version of the database. We isolated all human mature miRNAs in the database and treated them independently from their remaining species groups. Their average weighted clustering coefficients are shown in Fig. 6. Unlike the complete mature miRNA network, the clustering coefficient decreases from its initial version until version 6. From version 6 to 7 there is a steep increase and a peak, followed by a smooth decrease until the last

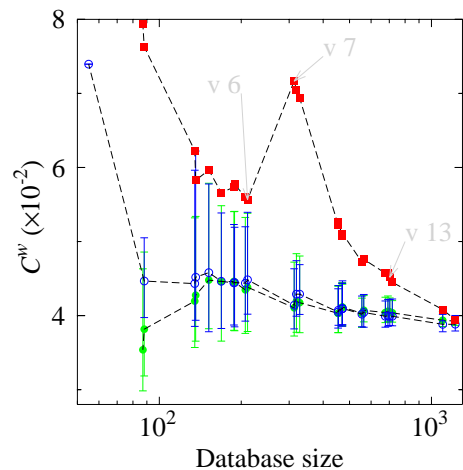


FIG. 6. Average clustering coefficients  $C$  of human mature miRNAs as a function of database size. Red solid boxes are for the real miRBase versions. The simulated database versions are represented by blue circles for a random initial database, and green boxes if they start from version 1.2.

version of miRBase under consideration. In this case we do not observe the sudden reduction between versions 13 and 14. While one may argue that there are only 15 new human miRNAs in version 14, no sudden drop is observed for any later versions even though there is an increase of nearly 400 new sequences between versions 14 and 15. It is interesting to note that the number of human miRNAs present in miRBase is already larger than some predicted upper estimate which were made recently [29]. Nevertheless, this means that the set of human miRNAs may be already quite complete. Yet even in this case, no simulation did reproduce the actual course of miRNA discovery. We have not attempted to calculate the network topology for other species groups in separate, as these groups are still too small to provide meaningful network topology parameters.

The results presented here let us to the conclusion that the sudden change in network topology between version 13 and 14 is not dominated by a change in similarity scores, although these scores do decrease continually since version 7. They are also not present in the largest species group. Clearly, there must have happened an important technological or methodological change in the way new miRNAs were discovered around the year 2009. Until 2005, the detection of miRNAs was usually preceded by an extensive bioinformatics analysis of known sequences. For example, some techniques would use conserved stem loops of precursor miRNAs [30, 31] or phylogenetic shadowing [32, 33] for identifying new miRNAs. These techniques rely heavily on sequence alignments of known miRNAs and this appears to be reflected by the steady increase of clustering coefficients up to 2005 seen in Figs. 1 and 5. For human miRNA the situation is less clear as shown in Fig. 6, there is an initial steady de-

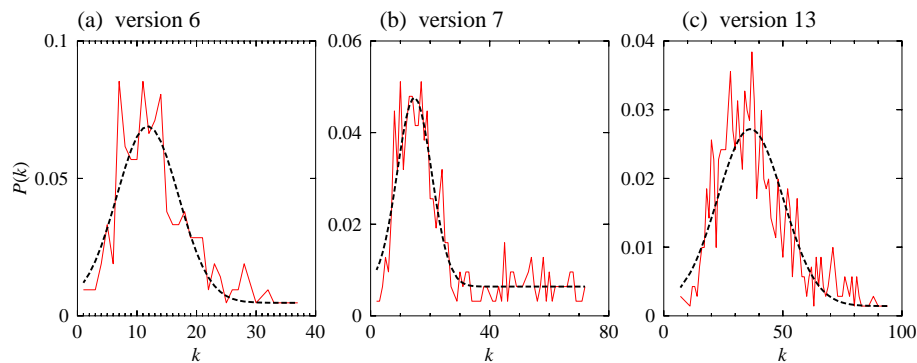


FIG. 7. **Connectivity distribution  $P(k)$  as a function of connections  $k$  of human mature miRNAs.** Shown are versions (a) 6, (b) 11 and (c) 16. Solid red curves are the  $P(k)$  distributions and dashed black curves are the calculated symmetrical Gaussian regression of  $P(k)$ .

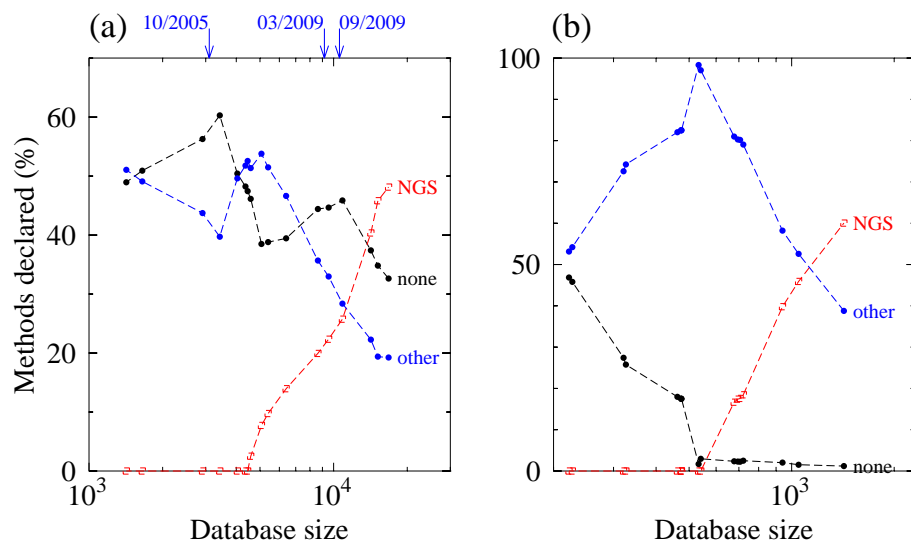


FIG. 8. **Number of experimental methods declared in miRBase.** Fractions are computed by counting the tags **experimental** declared for each miRNA. Tag values '464', 'SOLiD' and 'Solexa' are collectively shown as red boxes refer to next-generation sequencing (NGS) methods. Blue boxes are all methods different from NGS (other) and black bullets are sequences for which no method was given (none). Part (a) shows the complete database, and part (b) the human miRNAs only. Note that a large number of miRNAs are sequenced by several techniques, or resequenced with different techniques at a later time. Also, a large number of miRNAs have their experimental methods declared some time after their first appearance in the database. In other words, the annotation of methods is only approximately correlated with the appearance of new miRNAs in the database.

crease until version 6, then a steep increase for version 7. From 2005 onwards clustering coefficients decrease for all situations which coincides with the onset of RNA-seq, high throughput deep sequencing techniques widely used in studies of the transcriptome. This is shown in Figure 8 which presents the evolution of sequencing methods declared in the database over time. Differently from previous techniques, RNA-seq has some features that are responsible for the abundance of data generated by this technique: cDNA molecules are sequenced in parallel, producing large amounts of sequence data and the identification of sequences (like miRNAs) can be made without

prior sequencing knowledge [34]. In 2009 there appears to be deluge of RNA-seq data that completely changes the network topology. The computational methods to identify candidate miRNAs from the deep sequencing data operate under a very different set of requirements than earlier methods. In particular methods such as miRD-eep [35] explicitly avoid cross-species comparisons. Several more recent methods of miRNA discovery rely less on similarities with know miRNAs and employ other strategies such as biochemical characteristics of miRNA biogenesis [36] or try to predict Drosha processing sites to improve miRNA prediction [37]. This may explain

the steady decrease in clustering coefficients starting in 2005 and perhaps the sharp drop seen in 2009. For human miRNA on the other hand, where much more effort was spent in the early days of miRNA sequencing, the database appears to be much more complete and the drastic drop around version 13 and 14 is absent.

## CONCLUSIONS

We developed a procedure which makes use of sequence similarities to evaluate if network topologies could be biased by network growth. We applied the technique to the network topology analysis of the chronological history of sequences deposited in miRBase. We were able to show that the network topology, notably the clustering coefficients, shows a clear database construction bias. This means the resulting network topology depends critically at which point of time in the history of miRBase it was performed. For example, an analysis performed in

2009 would arrive at totally different network properties than the same analysis made a year earlier. We believe that this substantiates some of the criticism that indiscriminate interpretation of network topology has received recently [38, 39].

## AUTHOR CONTRIBUTION

GBS wrote the Perl scripts of the sampling simulations, CPSG performed the initial analysis on the network topology, DFL provided the biological analysis of the miRNA methods, ACS and GW provided conceptual advice and supervised GBS, CPSG and DFL, and wrote the paper.

## ACKNOWLEDGEMENTS

Funding: CNPq, Capes, Fapemig and National Institute for Science and Technology of Complex Systems.

- 
- [1] Barabási, A. and Albert, R. Emergence of scaling in random networks. *Science* **286**(5439), 509 (1999).
  - [2] Apicella, C., Marlowe, F., Fowler, J., and Christakis, N. Social networks and cooperation in hunter-gatherers. *Nature* **481**(7382), 497–501 (2012).
  - [3] Montoya, J., Pimm, S., and Solé, R. Ecological networks and their fragility. *Nature* **442**(7100), 259–264 (2006).
  - [4] Bascompte, J. Disentangling the web of life. *Science* **325**(5939), 416–419 (2009).
  - [5] Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* **101**(11), 3747–3752 (2004).
  - [6] Ahn, Y., Ahnert, S., Bagrow, J., and Barabási, A. Flavor network and the principles of food pairing. *Scientific reports* **1** (2011).
  - [7] Lesk, A. M. *Introduction to Bioinformatics*. Oxford University Press, (2008).
  - [8] Blow, N. Systems biology: untangling the protein web. *Nature* **460**(7253), 415–418 (2009).
  - [9] Lees, J., Heriche, J., Morilla, I., Ranea, J., and Orengo, C. Systematic computational prediction of protein interaction networks. *Physical Biology* **8**(3), 035008 (2011).
  - [10] Pan, Y., Durfee, T., Bockhorst, J., and Craven, M. Connecting quantitative regulatory-network models to the genome. *Bioinformatics* **23**(13), i367–i376 (2007).
  - [11] Tanaka, R., Yi, T.-M., and Doyle, J. Some protein interaction data do not exhibit power law statistics. *FEBS Letters* **579**(23), 5140–5144 (2005).
  - [12] Pržulj, N., Corneil, D. G., and Jurisica, I. Modeling interactome: scale-free or geometric? *Bioinformatics* **20**(18), 3508–3515 (2004).
  - [13] Khanin, R. and Wit, E. How scale-free are biological networks. *Journal of Computational Biology* **13**(3), 810–818 (2006).
  - [14] Stumpf, M., Wiuf, C., and May, R. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci. USA* **102**(12), 4221 (2005).
  - [15] Han, J., Dupuy, D., Bertin, N., Cusick, M., and Vidal, M. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature biotechnology* **23**(7), 839–844 (2005).
  - [16] Lee, S., Kim, P., and Jeong, H. Statistical properties of sampled networks. *Phys. Rev. E* **73**(1), 016102 (2006).
  - [17] Griffiths-Jones, S., Grocock, R., van Dongen, S., Bateman, A., and Enright, A. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140–D144 (2006).
  - [18] Kozomara, A. and Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucl. Acids. Res.* **39**(suppl 1), D152 (2011).
  - [19] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical Recipes in C*. Cambridge University Press, Cambridge, (1988).
  - [20] Weber, G., Essex, J. W., and Neylon, C. Probing the microscopic flexibility of DNA from melting temperatures. *Nature Physics* **5**, 769–773 (2009).
  - [21] Miele, V., Penel, S., and Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**(1), 116 (2011).
  - [22] Miele, V., Penel, S., Daubin, V., Picard, F., Kahn, D., and Duret, L. High-quality sequence clustering guided by network topology and multiple alignment likelihood. *Bioinformatics* **28**(8), 1078–1085 (2012).
  - [23] Andrade, R. F. S., et al. Detecting network communities: an application to phylogenetic analysis. *PLoS Computational Biology* **7**(5), e1001131 (2011).
  - [24] Atkinson, H., Morris, J., Ferrin, T., and Babbitt, P. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **4**(2), e4345 (2009).



- [25] Needleman, S. and Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**(3), 443–453 (1970).
- [26] Gonzalez, M. W. and Pearson, W. R. Homologous over-extension: a challenge for iterative similarity searches. *Nucl. Acids. Res.* **38**(7), 2177–2189 (2010).
- [27] Castro e Silva, A., Weber, G., Machado, R. F., Wanner, E. F., and Guerra-Sá, R. Identity transposon networks in *D. Melanogaster*. *Lecture Notes in Computer Science* **5167**, 161–164 (2008).
- [28] Onnela, J., Saramäki, J., Kertész, J., and Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E* **71**(6), 065103 (2005).
- [29] Pritchard, C., Cheng, H., and Tewari, M. MicroRNA profiling: approaches and considerations. *Nature Reviews Genetics* **13**(5), 358–369 (2012).
- [30] Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., and Bartel, D. P. Vertebrate microRNA genes. *Science* **299**(5612), 1540 (2003).
- [31] Lai, E., Tomancak, P., Williams, R., Rubin, G., et al. Computational identification of *drosophila* microRNA genes. *Genome Biol.* **4**(7), R42 (2003).
- [32] Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R., and Cuppen, E. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**(1), 21–24 (2005).
- [33] Berezikov, E., Cuppen, E., and Plasterk, R. H. A. Approaches to microRNA discovery. *Nat. Genet.* **38**(1), S2–S7 (2006).
- [34] Wang, Z., Gerstein, M., and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**(1), 57–63 (2009).
- [35] Friedländer, M., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. Discovering microRNAs from deep sequencing data using miRD-eep. *Nat. Biotechnol.* **26**(4), 407–415 (2008).
- [36] Hendrix, D., Levine, M., and Shi, W. Method miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol.* **11**, R39 (2010).
- [37] Helvik, S. A., Snøve, Ola, J., and Sætrom, P. Reliable prediction of Droscha processing sites improves microRNA gene prediction. *Bioinformatics* **23**(2), 142–149 (2007).
- [38] Hakes, L., Pinney, J., Robertson, D., and Lovell, S. Protein-protein interaction networks and biology-what’s the connection? *Nature Biotechnology* **26**(1), 69–72 (2008).
- [39] Lima-Mendez, G. and van Helden, J. The powerful law of the power law and other myths in network biology. *Molecular BioSystems* **5**(12), 1482–1493 (2009).